DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Land Use Area Classification Project

Submitted by

Shambhavi Kumari

Registration No:12322701

Programme and Section: CSE-2027 , K23GF

Course Code :INT375

Under the Guidance of

Aashima Maam

Discipline of CSE/IT

Lovely School of COMPUTER SCIENCE

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Shambhavi Kumari bearing Registration no 12322701 has completed INT375 project titled, **"Land Use Area Classification Project"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science
Lovely Professional University
Phagwara, Punjab.

Date: 08/04/2025

DECLARATION

I, Shambhavi Kumari, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 08/04/2025

Registration No:12322701 Shambhavi Kumari

Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the successful completion of this project.

First and foremost, I am deeply thankful to Aashima Maam my mentor and guide, for their constant support, valuable suggestions, and encouragement throughout the duration of this project. Their insights helped me explore the dataset more meaningfully and present my findings effectively.

I would also like to thank the Department of Computer Science and Engineering, LOVELY PROFESSIONAL UNIVERSITY for providing me with the necessary resources and a learning environment that made this project possible.

My heartfelt thanks to my friends and classmates who supported me during the various stages of this analysis, offering valuable feedback and motivation.

Lastly, I am grateful to the creators and maintainers of the dataset, as well as the open-source community for tools like Python, Pandas, Seaborn, and Matplotlib, which played an integral role in my data analysis journey.

This project has been a great learning experience, and I truly appreciate everyone who helped make it possible.

Table of Contents

Exploratory Data Analysis on Agricultural Land Usage in India

Introduction.

Exploratory Data Analysis (EDA) is a critical early step in the data science process. It helps understand the structure, trends, and patterns in the dataset before applying advanced techniques or models.

This report presents EDA on a dataset related to agricultural land usage in India. It includes histogram, bar, box, pie, scatter, and heatmap visualizations, as well as log transformations and outlier detection.

Source of Dataset

The dataset used in this project is sourced from the [Ministry of Agriculture and Farmers Welfare], Government of India. It contains state-wise agricultural data for different land use types such as:

- Net area sown
- Area under current fallows
- Net area cultivated
- Uncultivated area

The data is tabulated across various Indian states and provides a useful perspective for comparative and trend-based analysis.

Link:-https://ndap.niti.gov.in/dataset/7172

EDA Process

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, often using visual methods. This section describes the step-by-step EDA process followed in this project.

3.1 Importing Required Libraries

The analysis was performed using Python with the help of libraries such as:

- `pandas` – for data manipulation
- `numpy` – for numerical operations
- `matplotlib.pyplot` and `seaborn` – for visualization

3.2 Data Loading

The dataset was loaded using `pandas` from a `.csv` file. An initial overview was done using:

Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import seaborn as sns
df=pd.read_excel('File.xlsx')
df.head(1)
df.head()  # Displays the first 5 rows
df.info()  # Gives data types and null values
df.describe()  # Statistical summary
```

3.3 Data Cleaning

- Removed rows with null or missing values.
- Standardized column names (removed trailing spaces).
- Converted relevant columns to numeric data types.

3.4 Data Transformation

- Melted the DataFrame for advanced plotting (e.g., for box plots):
  `df_melted = df.melt(id_vars='srcStateName', var_name='Area Type', value_name='Area')`
- Log transformation was applied to columns with large numeric ranges to reduce skewness.
  `df['Net area sown_log'] = np.log1p(df['Net area sown'])`

3.5 Data Visualization Setup

- Subplots were used for comparative histogram and bar plot analysis.
- Visual styles were customized using Seaborn themes and custom color palettes.

4. Analysis on Dataset

4.1 Histogram Analysis

i. Introduction

Histograms are used to understand the distribution of numeric data by grouping it into bins. They help identify skewness, central tendencies, and spread.

ii. General Description

We plotted histograms of area-related attributes after applying logarithmic transformation to manage skewed distributions.

iii. Specific Requirements, Functions and Formulas

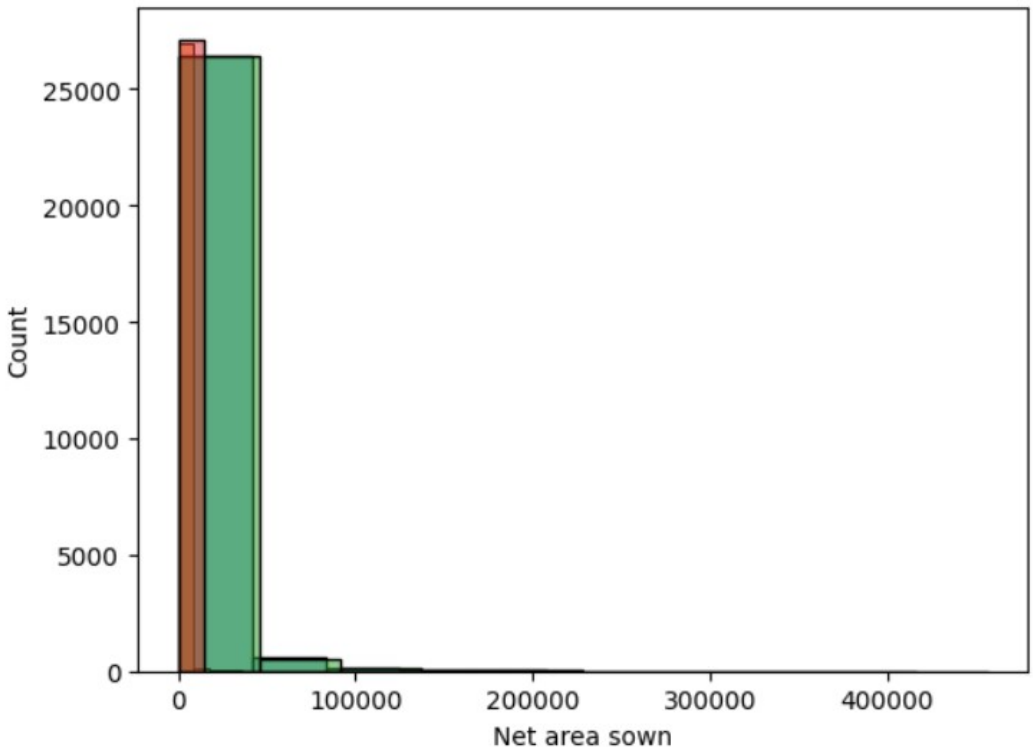
- Libraries Used: seaborn, matplotlib.pyplot
- Transformation Applied: np.log1p() to normalize skewed data.
- Function:
- sns.histplot(data=df, x='Net area sown_log', kde=True)
- **Grid of Subplots**: 2x2 layout using:

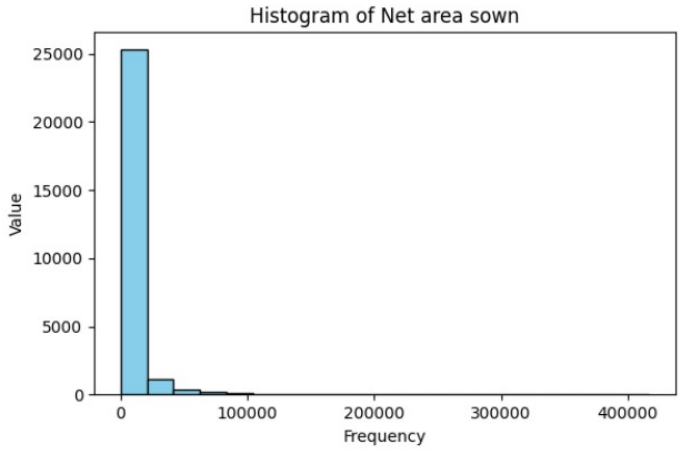
fig, axes = plt.subplots(2, 2, figsize=(12, 8))

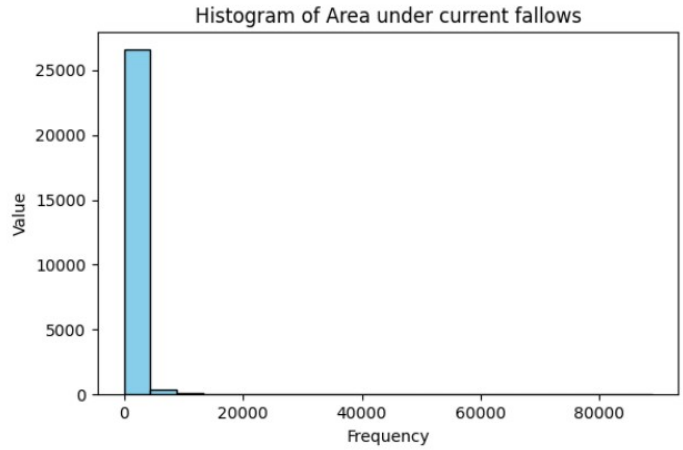
iv. Analysis Results

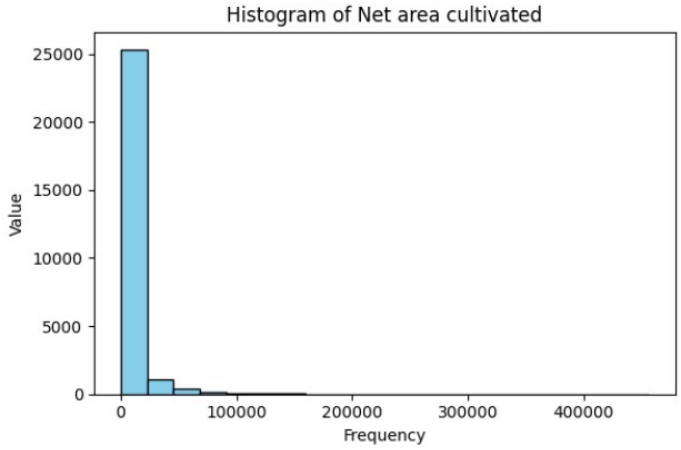
- The log-transformed histograms showed more balanced distributions.
- The KDE curves provided a clear view of the peak and spread.
- Some features still showed slight skewness, indicating natural distribution biases in state-wise land data.
- **v. Visualization**

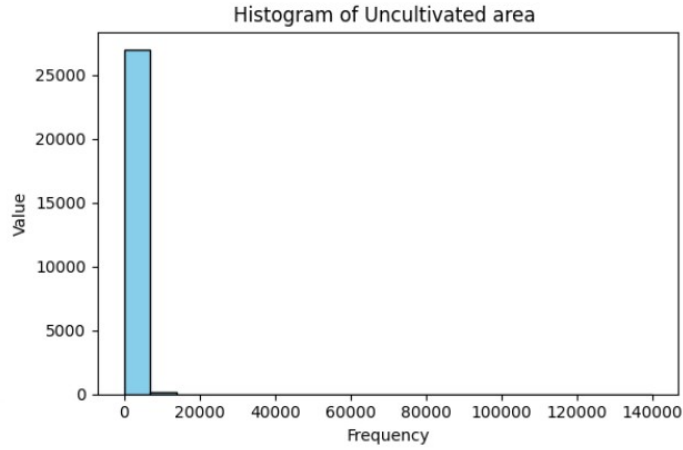
Log-Transformed Histogram of Selected Area Features

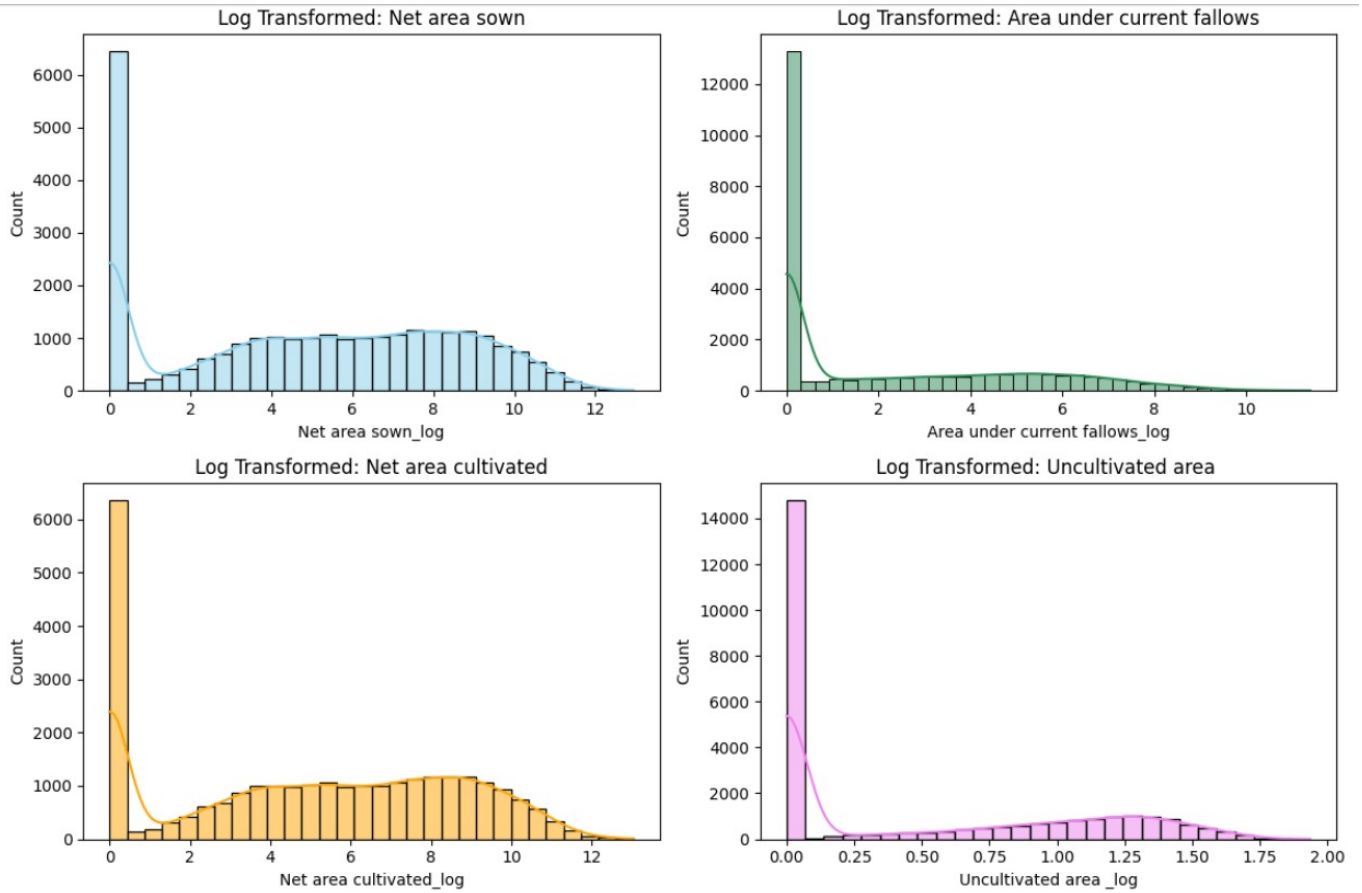



Histogram of Net area sown
Value
Frequency
0
5000
10000
15000
20000
25000
0
100000
200000
300000
400000


Histogram of Area under current fallows
Value
Frequency
0
5000
10000
15000
20000
25000
0
20000
40000
60000
80000


Histogram of Net area cultivated
Value
Frequency
0
5000
10000
15000
20000
25000
0
100000
200000
300000
400000


Histogram of Uncultivated area
Value
Frequency
0
5000
10000
15000
20000
25000
0
20000
40000
60000
80000
100000
120000
140000



4.2 Bar Plot Analysis

i. Introduction

Bar plots are useful for comparing numerical values across categorical groups. In this context, we visualized how different area types vary across states.

ii. General Description

Bar plots were created to compare total land features like *Net area sown*, *Net area cultivated*, *Area under current fallows*, and *Uncultivated area* across selected states.

iii. Specific Requirements, Functions and Formulas

- Data Aggregation: Grouped by srcStateName and used .sum()
- Function:

  sns.barplot(data=df, x='srcStateName', y='Net area sown')

- **Horizontal Bar Plot**: Used y for categorical axis and x for numerical

  sns.barplot(data=df, y='srcStateName', x='Net area sown')

- **Subplot Layout**: Created 2x2 grid of horizontal bar plots

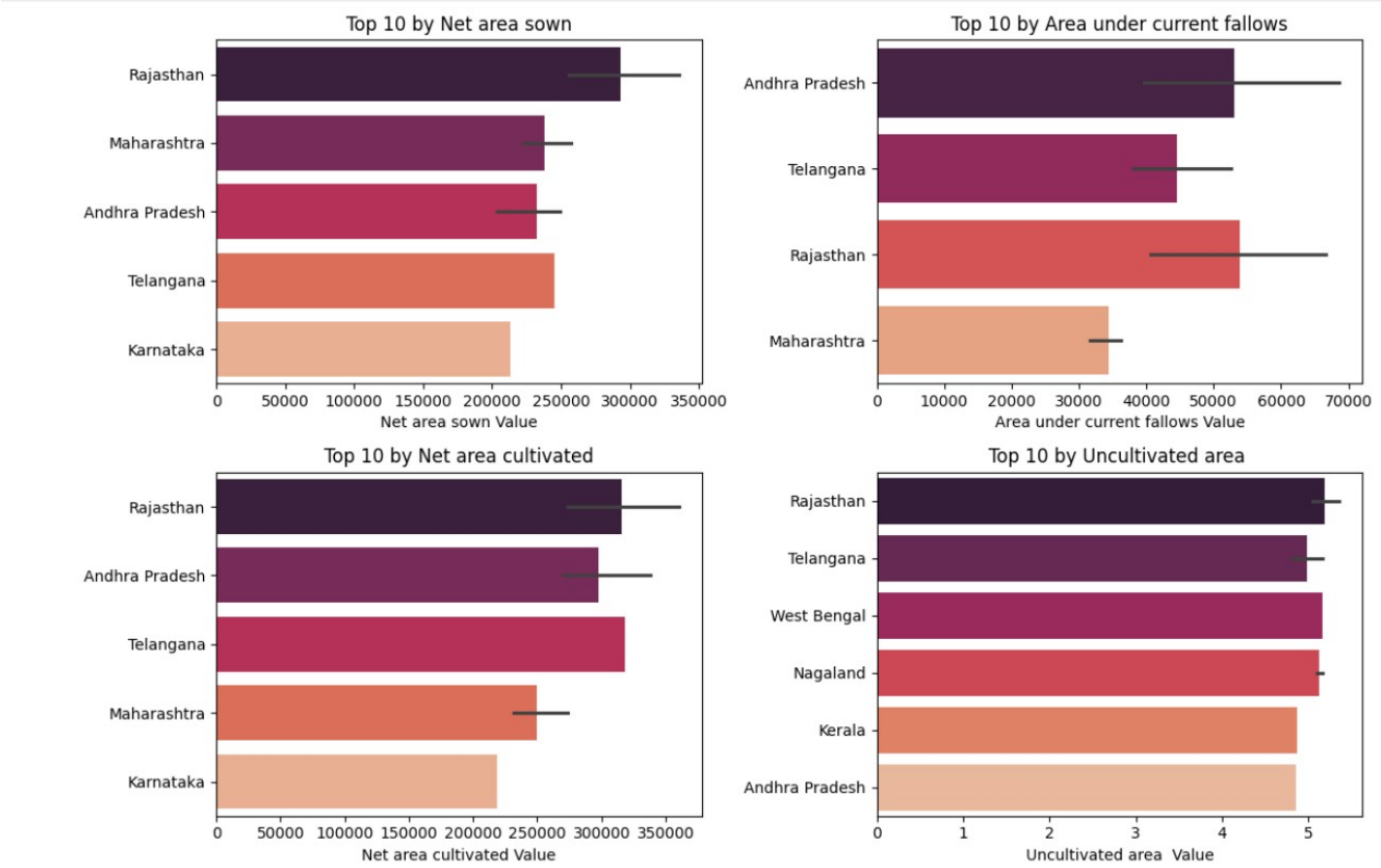
  fig, axes = plt.subplots(2, 2, figsize=(14, 10))

iv. Analysis Results

- States like Uttar Pradesh and Maharashtra showed higher values for net area sown and cultivated.
- Some states had larger uncultivated areas, indicating scope for agricultural expansion.
- Bar plots clearly showed the distribution gaps between states.

v. Visualization

- *Figure 4.2: Horizontal Bar Plots Comparing Different Area Types across States*



4.3 Box Plot Analysis

i. Introduction

Box plots are powerful tools to display the distribution, central tendency, and variability of numerical data. They also help in identifying outliers effectively.

ii. General Description

We used box plots to understand the distribution of different area types—like *Net area sown*, *Net area cultivated*, *Area under current fallows*, and *Uncultivated area*—across Indian states. The focus was on identifying outliers and skewness in the data.

iii. Specific Requirements, Functions and Formulas

- **Function Used:**

  sns.boxplot(x='Area Type', y='Area', data=df_melted)

- **Multiple Boxplots**: Used `melt()` to reshape the data from wide to long format

  df.melt(var_name='Area Type', value_name='Area')

- **Log Scale**:

  plt.yscale('log')

- **Grouped by State and Area Type**:

  sns.boxplot(x='srcStateName', y='Area', hue='Area Type', data=df_melted)

- **Filtered Top 10 States for Clarity**:

  top_states = df.groupby('srcStateName')['Net area sown'].mean().nlargest(10).index

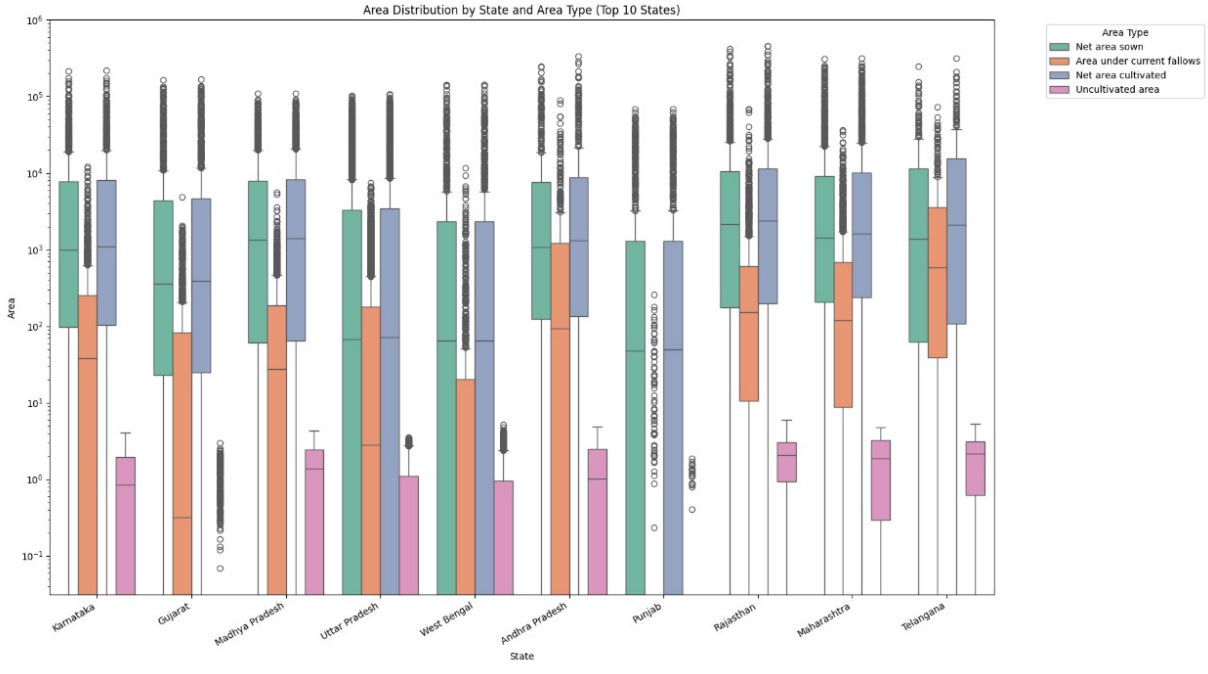
  df_filtered = df[df['srcStateName'].isin(top_states)]

  iv. Analysis Results

- All area types exhibited **left-skewed distributions**, suggesting a concentration of lower values.
- Several states had extreme outliers, especially in *Net area sown* and *Uncultivated area*.
- When grouped by state, box plots revealed the variation in land use patterns among top contributing states.

  v. Visualization

  Figure 4.3: Grouped Box Plot by State and Area Type



4.4 Pie and Donut Chart Analysis

i. Introduction

Pie and donut charts are useful for displaying the **proportional distribution** of categorical data. In this case, they help visualize how different area types contribute to total land usage.

ii. General Description

We created immersive pie and donut charts to represent the proportion of land categories such as *Net area sown*, *Area under current fallows*, *Net area cultivated*, and *Uncultivated area*. These charts give a clear, attractive snapshot of how land is utilized.

iii. Specific Requirements, Functions and Formulas

- **Pie Chart Function Used**:

  plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=140, shadow=True, explode=explode, colors=colors)

- **Donut Chart (Pie with Circle in Center)**:

  centre_circle = plt.Circle((0,0),0.70,fc='white')

  fig.gca().add_artist(centre_circle)

- **Enhancements Used**:
  - shadow=True for 3D effect
  - explode to highlight segments
  - Custom color palettes for visual appeal

iv. Analysis Results

- The **Net area sown** had the largest share in overall land use.
- **Uncultivated area** and **current fallows** accounted for smaller percentages.
- Both charts helped summarize categorical data at a glance, useful in presentations and reports.

v. Visualization

Figure 4.4: Immersive Pie Chart of Area Types

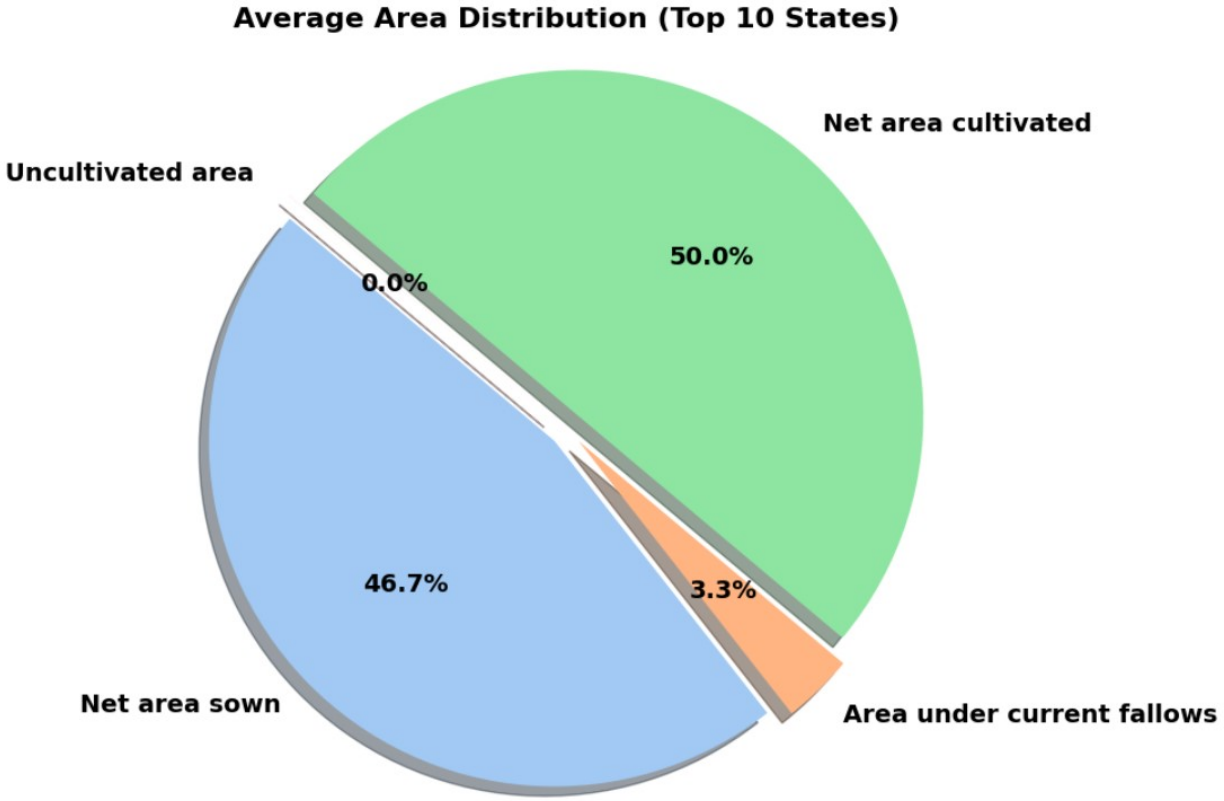
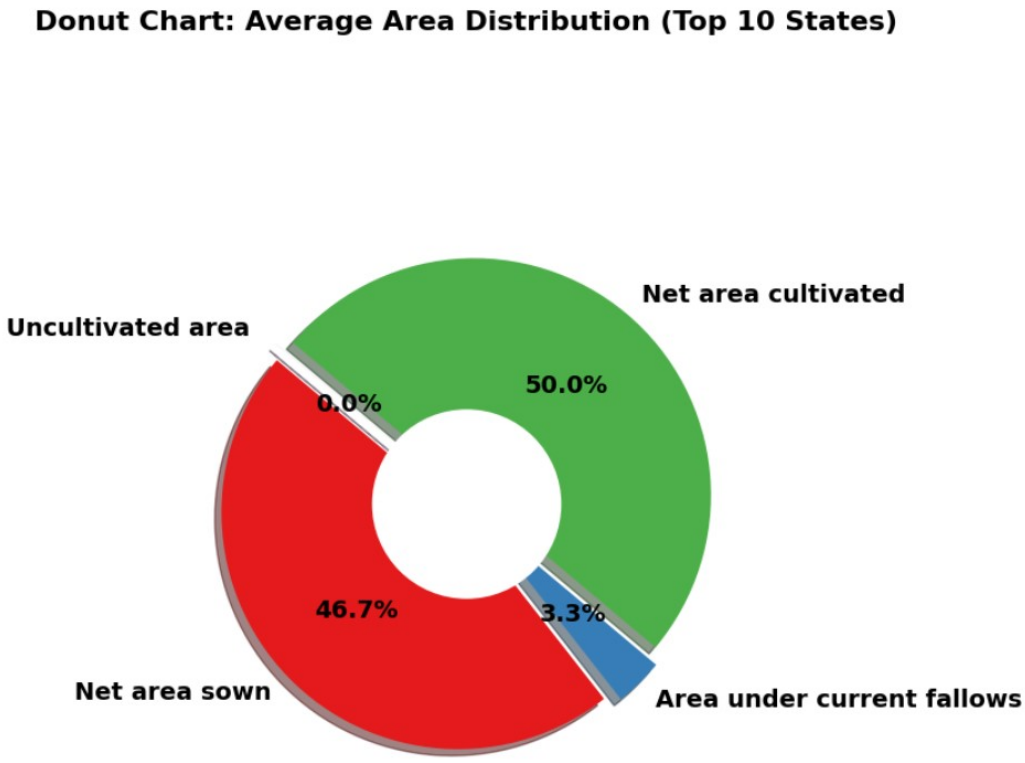


Figure 4.4.1: Immersive Donut Chart of Area Types



4.5 Scatter Plot Analysis

i. Introduction

Scatter plots are used to display relationships between two continuous variables. In this analysis, we use scatter plots to examine the correlation between various land use features.

ii. General Description

Scatter plots help identify trends, clusters, and outliers by plotting data points for each observation. This visualization assists in checking how one area type may influence another.

iii. Specific Requirements, Functions and Formulas

- **Function Used**:

  sns.scatterplot(data=df, x='Net area sown', y='Net area cultivated', hue='srcStateName', palette='Set2', s=100)

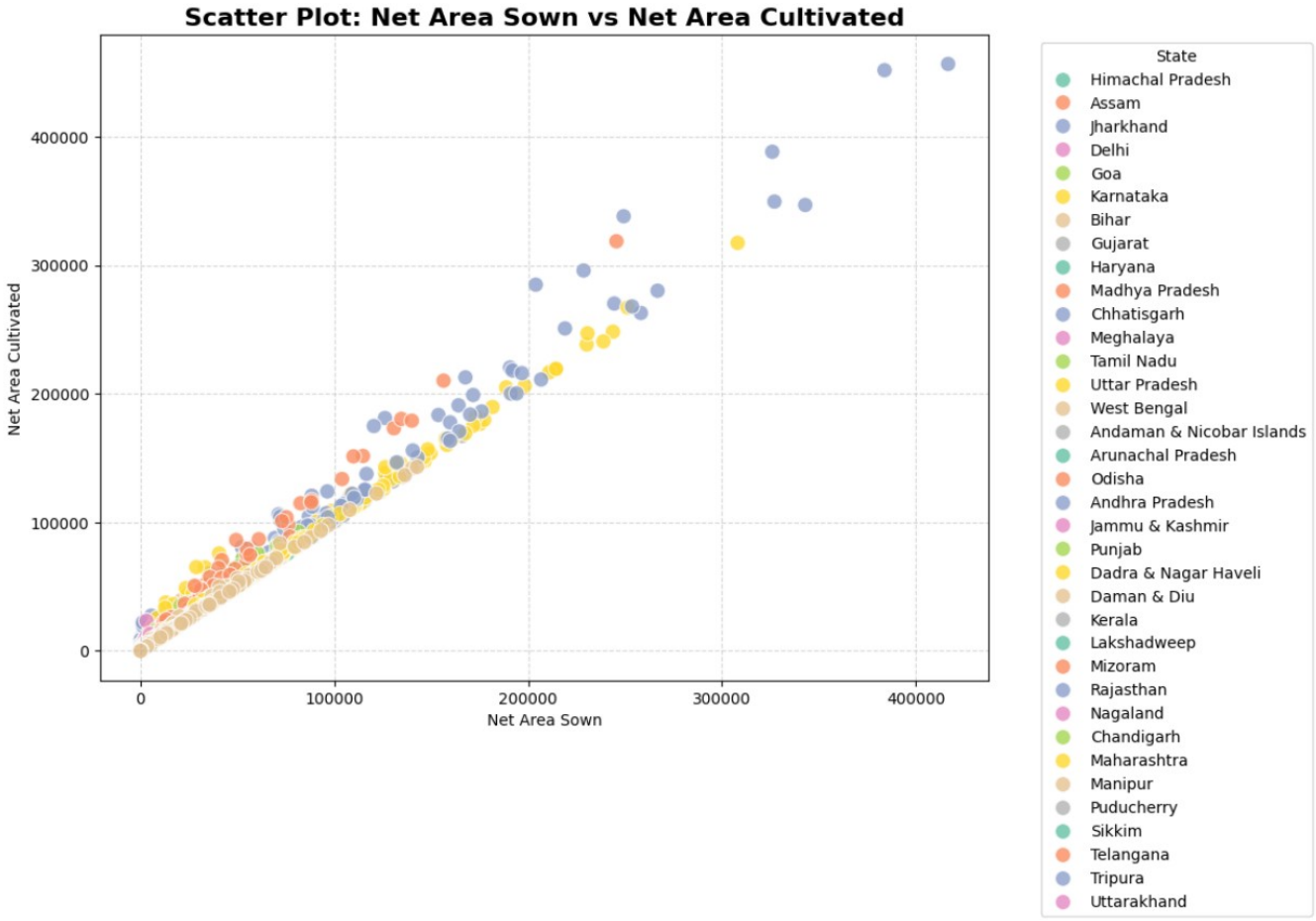
- **Enhancements**:
  - hue='srcStateName' distinguishes states using colors.
  - s=100 makes each point large and more visible.
  - Custom palette (Set2) provides a colorful and clear display.
- **Optional Addition**: plt.xscale('log') or plt.yscale('log') if data ranges widely.

iv. Analysis Results

- A **positive correlation** was observed between *Net area sown* and *Net area cultivated*.
- Some states had unusually high values, hinting at either better agricultural infrastructure or reporting differences.
- This plot helped highlight the comparative performance of states in utilizing cultivated land.

v. Visualization

Figure 4.5: Scatter Plot – Net Area Sown vs Net Area Cultivated by State



4.6 Heatmap Analysis

i. Introduction

Heatmaps are an effective way to visualize the **correlation** between numerical features in a dataset. They allow us to quickly identify which variables have strong positive or negative relationships.

ii. General Description

A heatmap uses color gradients to show the degree of correlation between different columns. It is particularly helpful for **feature selection**, **pattern recognition**, and spotting **redundant variables** in a dataset.

iii. Specific Requirements, Functions and Formulas

- **Function Used**:

  correlation = df[columns].corr()

  plt.figure(figsize=(10, 8))

  sns.heatmap(correlation, annot=True, cmap='YlGnBu', fmt=".2f", linewidths=0.5, linecolor='white')

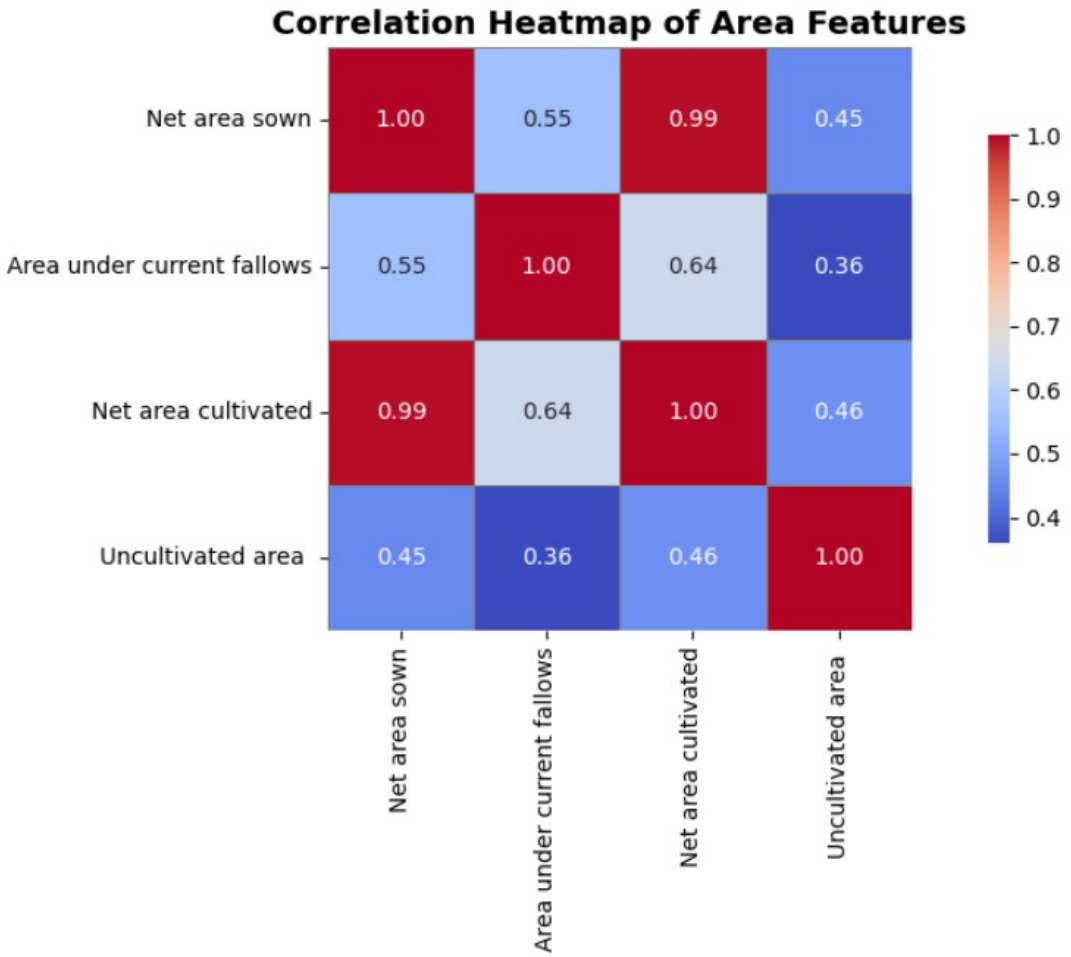
- **Explanation**:
  - df[columns].corr() calculates the correlation matrix.
  - annot=True displays the correlation values in each cell.
  - cmap='YlGnBu' gives a blue-green color gradient for clarity.
  - linewidths and linecolor enhance readability.

iv. Analysis Results

- High correlation (close to +1) was found between *Net area sown* and *Net area cultivated*, which is expected.
- Other area types like *Area under fallows* showed weaker or even negative correlations with cultivated areas.
- The heatmap helped identify **which columns move together** and which are independent.

v. Visualization

Figure 4.6: Heatmap Showing Correlation Between Area Features



5. Conclusion

The Exploratory Data Analysis (EDA) conducted on the agricultural dataset provided valuable insights into land usage patterns across Indian states. Various visualization techniques such as **histograms, bar plots, box plots, pie charts, donut charts, scatter plots, and heatmaps** were used to understand the distribution and relationships among features like *Net area sown*, *Uncultivated area*, *Net area cultivated*, and *Area under current fallows*.

Key findings include:

- Most area-related variables were **left-skewed**, indicating a concentration of values at the higher end.
- Significant **variation across states** was observed in terms of area sown and cultivated.
- Certain variables such as *Net area sown* and *Net area cultivated* had strong **positive correlations**, as expected.
- Visualizing outliers using **boxplots** and **heatmaps** enabled a deeper understanding of data irregularities and dependencies.

Through this process, we gained a holistic view of how agricultural land is being used in different regions, which could serve as a reference point for future agricultural planning and development.

6. Future Scope

The current analysis provides foundational insights into agricultural land usage; however, it opens several avenues for future exploration:

- **Inclusion of Temporal Data**: Adding data across multiple years could help observe trends over time and understand the effects of policy changes or climate events.
- **Machine Learning Models**: Predictive modeling can be applied to forecast agricultural patterns, crop yields, or land usage efficiency using regression and classification techniques.
- **Integration with Climate Data**: Correlating land usage with rainfall, temperature, and soil quality would enhance decision-making in sustainable farming.
- **State-level Recommendations**: A deeper drill-down into state-wise performance could help governments and NGOs tailor solutions for land development and resource management.
- **Geospatial Visualization**: Using tools like Folium or GIS-based mapping could provide spatial insights and make the data more intuitive and actionable.

This future work would enhance the usability of the dataset for agricultural planning, policymaking, and academic research.

7. References

[1] Ministry of Agriculture and Farmers Welfare, Government of India, *Agricultural Statistics at a Glance*, [Online]. Available: https://agricoop.nic.in/

[2] Python Software Foundation, “Python Language Reference, version 3.10,” [Online]. Available: https://www.python.org

[3] Wes McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed., O’Reilly Media, 2017.

[4] Michael Waskom, “Seaborn: Statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, pp. 3021, 2021. [Online]. Available: https://seaborn.pydata.org/

[5] Hunter, J. D., “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[6] Jupyter Project, “Project Jupyter,” [Online]. Available: https://jupyter.org/

[7] T. E. Oliphant, “A Guide to NumPy,” USA: Trelgol Publishing, 2006.